CA Journal of Cultural Analytics

The Migrant Letter Digitised: Visualising Metadata

Niall O'Leary, Emma Moreton

03.28.17

Peer-Reviewed By: Ian Milligan

Clusters: Correspondence

Article DOI: 10.22148/16.013

Dataverse DOI: 10.7910/DVN/ERQWPH

PDF: 10.31235/osf.io/89xbp

Journal ISSN: 2371-4549

Cite: Niall O'Leary and Emma Moreton, "The Migrant Letter Digitised: Visualising Metadata," Journal of Cultural Analytics. March 28, 2017. DOI: 10.31235/osf.io/89xbp

Within the digital humanities, social network analysis - using digital technologies to examine the relationship between people, places and things - has explored a wide range of digital communication formats, from emails to tweets.[1] This has been made possible because of the large amount of online digital data and has spawned many new techniques specifically aimed at analysing very large datasets, often termed Big Data. The quantity of data resulting from digital communication is enormous, and therefore a tempting source of raw material. However, there is also a long tradition of non-digital communication, letter-writing, which shares many of the formal characteristics of digital formats and also constitutes a huge body of data.

[1] For an example see: Martin Grandjean, "A social network analysis of Twitter: Mapping the digital humanities community," *Cogent Arts & Humanities* 3(1), (2016): 1171458. Accessed March 15, 2017.

This article builds on workshop discussions that took place as part of an AHRC funded research networking project.[2] The project, *Digitising experiences of migration: the development of interconnected letters collections* (DEM), started in April 2013 and finished in July 2014.[3] Through a series of workshops, the project brought together scholars from different disciplines currently working with migrant letters as a primary data source, to explore the digital potential of these perhaps iconic documents. The workshops looked at how letters are being used across the disciplines, identifying similarities and differences in transcription, digitisation and annotation practices. The aims were to examine issues and challenges surrounding digitisation, build capacity relating to correspondence mark-up, and initiate the process of interconnecting resources to encourage cross-disciplinary research. Subsequent to the DEM project, TEI templates were developed for capturing information within and about migrant correspondence, and visualisation tools were trialled with metadata from a sample of letter collections.[4] Additionally, as a demonstration of how the project's outputs could be repurposed and expanded, the correspondence metadata that was collected for DEM was added to a more general correspondence project, *Visual Correspondence*.[5] This article reports on some of this subsequent work. It will discuss some of the challenges and opportunities of interconnecting correspondence collections and how visualisation tools may be used to explore the metadata of letters. While many of the lessons learnt apply to correspondence in general, the emphasis is on the migrant letter.

Online Letters - Projects and Opportunities

There are some barriers to the large-scale analysis of letters, the most obvious being their physical nature. To use digital techniques to analyse them, letters, or at least metadata about letters, must be digitised. This can be a laborious, not to mention expensive, undertaking. Nevertheless, many correspondence collections have been digitised and are available online - primarily, though not exclusively, those of well-known historical figures. Well-known collections include

[2] This work was supported by the Arts and Humanities Research Council [AH/K006231/1]

[3] Coventry University 2013-2014 *Digitising Experiences of Migration: The Development of Interconnected Letter Collections*. Available from: http://lettersofmigration.blogspot.com. Accessed March 15, 2017.

[4] An initial set of visualisations created for the DEM project and based on two datasets is available from: http://www.nialloleary.ie/development/correspondence/correspondence.php. Accessed March 15, 2017.

[5] *Visual Correspondence*. Available from: http://www.correspondence.ie./. Accessed March 15, 2017.

the *Mark Twain Online Project*,[6] the *Darwin Correspondence Project*,[7] and the *Alfred Russel Wallace Correspondence Project*.[8] These projects are mostly concerned with the content of individual letters and few attempt to visualise their collections. However, the *Letters of the 1916* initiative tries to augment its collection of crowdsourced correspondence by tagging the themes present in its letters.[9] Focusing more on the linguistic aspects of letter content, Chris Culy, formerly of University of Tübingen, offers an innovative suite of tools for performing linguistic analyses of correspondence generally, and provides some examples using letters from Barratt/Browning, Michelangelo and Ambrose Bierce.[10] These broader approaches move beyond the individual letter to a more holistic perspective.

Large though any one of these collections might appear (the Mark Twain collection numbers some 30,000 records), in Big Data terms, which often deals with tens of thousands to millions of records, few are suitable for analysis in themselves. To apply large-scale analytical techniques, ideally, collections must be aggregated. However, off and online projects concerned with letters are often developed in isolation. As a group they lack consistency in terms of objective and often use a variety of formats to publish their data. All they seem to have in common are their raw material, the letters themselves. If social network analysis is to be brought to bear on correspondence we need to aggregate letters in a new way.

Metadata provides a possible solution. The web service *correspSearch,*[11] for instance brings together letters from several collections, such as *Carl Maria von Weber - Collected Works* and *Intellectual Berlin around 1800*, using data drawn from TEI headers. Stanford's *Mapping the Republic of Letters*[12] has also indexed a variety of sources of 16th, 17th and 18th century letters for its site. In addition, it has augmented this index with a range of visualisations that explore the metadata of the correspondence, albeit in a somewhat fragmented way. Oxford's *Early Modern Letters Online*[13] is also a part of this project and catalogues 95,000 letters

[6] *Mark Twain Project Online*. Available from: http://www.marktwainproject.org/.

[7] University of Cambridge 2015. *Darwin Correspondence Project*. Available from: http://www.darwinproject.ac.uk/. Accessed March 15, 2017.

[8] Wallace Correspondence Project 2013. *The Alfred Russel Wallace Correspondence Project*. Available from: http://wallaceletters.info/content/homepage. Accessed March 15, 2017.

[9] National University of Ireland Maynooth 2015. *Letters of 1916* [Online]. Available from: http://letters1916.maynoothuniversity.ie/. Accessed March 15, 2017.

[10] Culy, C. 2010. *Visualization of Language and Linguistic Data*. Available from: http://chrisculy.net/. Accessed March 15, 2017.

[11] Berlin Brandenburg Academy of Sciences and Humanities 2014. *correspSearch*. Available from: http://correspsearch.bbaw.de/. Accessed March 15, 2017.

[12] Mapping the Republic of Letters 2013. *Mapping the Republic of Letters*. Available from: http://republicofletters.stanford.edu/. Accessed March 15, 2017.

[13] Cultures of Knowledge Project 2009. *Early Modern Letters Online*. Available from: http://emlo.

from the early modern era. It acts as an umbrella for a wide variety of smaller projects, such as the *Women's Early Modern Letters Online* project.[14] There is, then, some precedent for bringing together smaller projects for the purposes of analysis and for exploring such data through visualisation. Generally, though, historical migrant correspondence offers a relatively unexploited source of Big Data ripe for analysis.

As with more general correspondence projects, migrant correspondence projects have often evolved independently of one another with varying aims and approaches. Important studies of, for example, English, Scottish, Welsh, Irish, German, Swedish and Norwegian migrants have demonstrated the value of using personal letters to gain a fuller and deeper understanding of both the complex social processes of migration and the conditions and daily lives of the migrants themselves.[15] However, while research teams have been successful in tackling important questions relating to social history and migration studies, relatively few projects have moved beyond the digitisation stage to enhance usability and searchability through the use of digital technologies. Different migrant letter collections cannot easily interconnect if they are simply digitised without some semantic and structural tagging. Indeed, even where mark-up has been added, for instance using TEI, it may concentrate on elements only relevant to a particular project. For instance mark-up concentrating on linguistic forms does not lend itself easily to an analysis of socio-economic factors. If we are to aggregate migrant letters in a similar manner to *correspDesc* or the *Republic of Letters* we need to be able to aggregate disparate collections based on elements

bodleian.ox.ac.uk/. Accessed March 15, 2017.

[14] Women's Early Modern Letters Online [WEMLO] 2016. *Women's Early Modern Letters Online*. Available from: http://emlo-portal.bodleian.ox.ac.uk/collections/?page_id=2595. Accessed March 15, 2017.

[15] See, for example, David A. Gerber, *Authors of Their Lives: The Personal Correspondence of British Immigrants to North America in the Nineteenth Century* (New York: New York University Press, 2006); Charlotte Erickson, *Invisible Immigrants: The Adaptation of English and Scottish Immigrants in Nineteenth-Centrury America* (London: Weidenfeld & Nicolson, 1972); Alan Conway, *The Welsh in America: Letters from the Immigrants* (Minneapolis: University of Minnesota Press, 1961); Arnold Schrier, *Ireland and the Irish Emigration, 1850-1900* (Minneapolis: University of Minnesota Press, 1958); Kerby A. Miller, *Ireland and Irish America: Culture, Class, and Transatlantic Migration* (Dublin: Field Day, 2008), Kerby A. Miller, *Migrants and Exiles: Ireland and the Irish Exodus to North America* (New York: Oxford University Press, 1988); Kerby A. Miller, Arnold Schrier, Bruce D. Boling, and David N. Doyle, *Irish Immigrants in the Land of Canaan: Letters and Memoirs from Colonial and Revolutionary America, 1675-1815* (New York: Oxford University Press, 2003); David Fitzpatrick, *Oceans of Consolation: Personal Accounts of Irish Migration to Australia* (Cork: Cork University Press, 1994); W. D. Kamphoefner, Wolfgang Helbich, and Ulrike Sommer, *News from the Land of Freedom: German Immigrants Write Home* (Ithaca: Cornell University Press, 1988); H. Arnold Barton, *Letters from the Promised Land: Swedes in America, 1840-1914* (Minneapolis: University of Minnesota Press, 1990); and Solveig Zempel, *In Their Own Words: Letters from Norwegian Immigrants* (Minneapolis: University of Minnesota Press, 1991).

that they share.

Fortunately, letters share some common attributes. At a basic level there is often (although not always) a sender, a recipient, an origin, a destination and a date. Bringing together these five elements from different migrant letter corpora allows their metadata to be compared and contrasted. This can provide a macro analysis of the letter writers and the communities within which they wrote. By concentrating on this metadata it is possible to aggregate the output of several projects regardless of their individual aims in order to attain the numbers of items needed for employing Big Data techniques of analysis. Central to such techniques is data visualisation. Data visualisations allow the user to see the grand patterns and the unique outliers pertaining to large quantities of data. In this respect, they help to make sense of Big Data and present alternative methods of searching and navigating huge datasets. Rather than an alternative to close reading, data visualisation is, we want to argue, a necessary complement.

Challenges

For this study we focused on four letter collections in particular: the *Irish Emigration Database* (IED), the Lough family letters, the Smith/O'Gowan correspondence, and the *Digitzing Immigrant Letters* collection. The IED is housed at the Mellon Centre for Migration Studies at the Ulster American Folk Park Museum in Northern Ireland, and contains over 4,000 letters by Irish migrants and their families, dating from the seventeenth to twentieth century.[16] The Lough family letters, part of Professor Kerby Miller's archive of approximately 5,000 Irish migrant correspondence, is housed at the University of Missouri and contains 99 letters dating from 1876 to 1928. The Smith/O'Gowan correspondence is a private collection donated by Michael Handford and contains 18 letters between Irish migrant Hugh Smith and his brother James dating from 1887 to 1908.[17] The

[16] The collaborative *Documenting Ireland: Parliament, People and Migration* (DIPPAM) project is an online archive of sources hosted by Queen's University, Belfast, that relates to the history of Ireland and the migratory experiences of its population between 1700 and the twentieth century. It consists of three principal databases: (a) Enhanced British Parliamentary Papers on Ireland (EPPI); (b) Irish Emigration Database (IED) referred to here, and (c) Voices of Migration and Return (VMR).

[17] What follows is some brief background information relating to the Smith/O'Gowan letters provided by the donor, Michael Handford. The family name is Smith, which the family changed from O'Gowan, during the time of Cromwell. The Smith/O'Gowan family moved to Drumbrucklis, County Cavan from somewhere in north Ulster when Cromwell was enforcing the Act of Settlement. According to family legend, they chose Drumbricklis because it was so remote (they thought Cromwell's troops would not find them there). Kathleen Smith migrated to Fall River Massachusetts to live with

Digitzing Immigrant Letters (DIL) collection, housed at the Immigration History Research Centre at the University of Minnesota, contains 85 letters by migrants and their families mainly in Europe and North America.[18] These collections were particularly well suited to the aims of the DEM project because of their variability in terms of size, accessibility and metadata, helping us to understand best practices across a range of different collections that are representative of the type of material that is currently available. Initial work concentrated on the IED and DIL collections. The Lough and Smith/O'Gowan datasets were added when the data was imported into the *Visual Correspondence* site.

Obviously, standardising metadata relating to several thousand letters is a major challenge. As described above, we decided to focus on five information categories that appeared to be common to all the letters: sender, recipient, origin, destination, and date. Although TEI was not used to tag all letters (time and resources did not allow this), where it was used the sender and recipient names were captured with the <correspAction> element and the @type attribute. Origin and destination were captured using the <settlement> element and the @key attribute. Finally, date information was captured using the <date> element and the @when attribute.

Metadata relating to names, places and dates are common across letter collections, reflecting some of the formal characteristics of correspondence in general. This means that the letters in a large dataset such as the IED can be compared and contrasted with smaller collections such as the Lough and Smith/O'Gowan letters. However, while it is true that formal similarities make all letters theoretically comparable, there are some caveats that need to be made. Conventions for representing names, places, and dates can vary from collection to collection with local conventions often being adopted instead of accepted global standards. In short, one cannot naively take the metadata from two distinct projects and compare their letters. Taking names as an example, while the 'surname/forename(s)' convention might seem standard, and indeed makes the alphabetical sorting of data particularly effective, it is not always used or indeed useful. In the case of Spanish and Portuguese names, for instance, the surname is not always easily identifiable; naming conventions have changed in Spanish/Portuguese speaking countries over the centuries and can encode information such as birthplace, ma-

her uncle - James Smith. James was in the police force and already had two daughters. All three girls became teachers in Massachusetts. None of them married, and they all lived together in the same house until they died. All but one of the Smith/O'Gowan letters were sent to that home, either from Drumbrucklis or from James's brother in New York. After the women's deaths, the letters were discovered when the house was cleared out and the correspondence was passed to relatives living at the family farm in Drumbrucklis.

[18] *Digitizing Immigrant Letters*. Available from: https://www.lib.umn.edu/ihrca/dil. Accessed March 15, 2017.

ternal heritage, and other elements, depending on the era. The name 'Ana Maria do Espírito Santo', for example, includes a nickname indicating no known father. Place names too have changed over the centuries. Some places have changed their spelling, but other locations have vanished or been incorporated into larger towns or even different countries. On the positive side, once a place has been identified and its geo-spatial coordinates recorded, it can be reconciled with other variants. Finally, the issue of dates is particularly problematic. The United States tends to use the convention 'MM-DD-YYYY' while most of the rest of the world use 'DD-MM-YYYY'. Obviously some reformatting is required before any proper comparison can be made; even the use of hyphens as opposed to slashes or dots can make comparing dates difficult.

Aside from these editorial issues, the letters themselves often pose the most challenging issues. Correspondents did not always include a year, month or day on their letters meaning that date information is often partial or absent. Many of the Lough letters, for example, are undated but their content has allowed them to be placed within an approximate timeframe. In some cases that timeframe has been narrowed down to a specific year; in other cases it has been narrowed down to within a ten-year span. While it is possible to capture a specific year as a separate field, recording date ranges is problematic. Databases offer a date field - a data type specifically geared to dates - which provides a range of computational functionality when populated with a date/time value. However, such a field cannot accommodate a date range and even if it could, writing programs to compare one date range with another, or even a single date, is awkward. Nevertheless, despite these challenges, any collection of letters from any available project or publication can be compared, contrasted and amalgamated, if its metadata can be captured in a central system.

The four sources used for this trial study were varied: while the IED and DIL collections are available online, the Lough and Smith/O'Gowan collections are offline resources and had to be manually transcribed and tagged in TEI. OxGarage was used to extract the metadata from TEI into a CSV format where applicable. In other cases, spreadsheets were compiled by hand or extracted from a database. The metadata was harvested on a project by project basis, with the idiosyncratic approaches of each collection identified and normalised as encountered. As the DEM project looked at the use of TEI in correspondence studies, datasets were initially stored as XML files. Extensible Stylesheet Language (XSL) transformations were used to convert them into a format suitable for online representation. XML is particularly good when dealing with semi-structured data, but as our data became cleaner and more structured a database became more appropriate. Databases can also provide a lot of tools that make development easier

(text searching, for example). For ease of development, robustness and storage the datasets were imported into a database, specifically the open-source MySQL.

By the time the Lough and Smith/O'Gowan collections were incorporated, tools were developed (using PHP, JavaScript and a backend MySQL database) to aid in the normalisation of data. To solve some of the issues related to names for instance, bespoke software was developed to parse names into a 'surname, forename' format. Although this format worked for the majority of cases, as has been described, conventions do vary and each name needed to be examined. A semi-automatic solution was developed whereby names and their standardised versions could be displayed on a web form alongside tick boxes. This web page allowed each name to be reviewed quickly and approved by ticking each box and submitting the form. Places too were of minimal use without geo-spatial coordinates. Software was developed, using SPARQL Protocol and RDF Query Language (SPARQL), for the retrieval of geo-coded results. This was good at suggesting possible coordinates, but the data could never be naively trusted - for instance a search for Venice might bring up results for both 'Venice, Italy' and 'Venice, California'. Again a web interface allowed a reviewer to review the suggestions, approve accurate location data and submit it to a MySQL database. All other locations (the majority) were manually identified using Google Maps and submitted to the database, again using custom-built web-based software. Fortunately once a location was identified it could be applied to all letters that used it. An address such as 'New York' was used again and again. Unfortunately differences in spelling or even the granularity of the address (e.g. 'New York, New York'), meant the majority of place names were not common across collections and had to be manually collected.

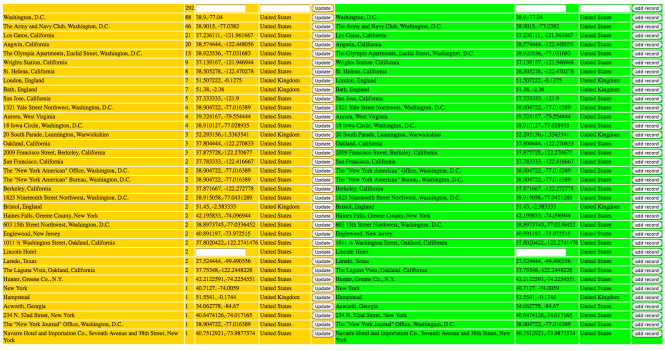

Figure 1. Web tool for entering geo-spatial coordinates.

From a programming perspective, the work described so far was a very labour

intensive process. Reviewing metadata and collecting geospatial coordinates was also extremely labour intensive. Although we are only dealing with a small subset of the metadata - sender, recipient, origin, destination, and date - it was essential that each element be normalised and geo-spatial coordinates captured separately for each location. Once stored in a database in a normalised way, these distinct projects with their distinct objectives - the (online) IED and DIL resources and the (offline) Lough and Smith/O'Gowan collections - could be treated as one amalgamated dataset. In the end, the metadata for each letter conformed to a standardised format allowing the database to be queried in a consistent manner and the results visualised.

Visualising metadata

The final database formed part of a web-based solution that used a server-side program (written in PHP) to query data and present it to the user in a web browser. When the database was queried the results were formatted into JSON (JavaScript Object Notation), a lightweight data-interchange format that is particularly well suited to JavaScript programming. This was delivered together with dynamically generated web pages that used JavaScript to visualise the data. For the visualisations themselves a variety of JavasScript libraries were used, but were tailored specifically for spatial, temporal and personal attributes.[19] Although visualisations use correspondence metadata from thousands of letters, paradoxically they reveal a lot about individual letter writers. In particular, they provide insights into three main areas:

1. *Activity*

A person's letters can be used to track their movements over time. While a close reading of a particular corpus may yield this narrative, a diagram generated automatically from the accumulated letters' metadata is far quicker to create and more effective in its descriptive power. Not only that, but because the generation of such graphs is interactive we can query the data from a variety of perspectives and pursue particular lines of enquiry without having to retrace our steps. By inspecting the frequency of letters, a sense of the correspondents' activities and personal relationships can be discerned, providing a vibrant picture of their lived experiences.

[19] These open-source libraries included Leaflet (http://leafletjs.com/), Exhibit (http://www.simile-widgets.org/exhibit3/), JQuery (http://jquery.com/), D3 (https://d3js.org/), Dimple (http://dimplejs.org/), and Sigmajs (http://sigmajs.org/).

2. *Connections*

Using the sender and recipient information, a vision of the correspondents' social network can be composed. By looking at each node in the social web, a sense of closeness or distance between participants can be seen. Indeed, this may in itself suggest further questions: why might one person not have been in communication with another, for instance?

3. *Unanswered Questions*

For many of the letter writers studied in this essay there are biographical blanks in what is known about them. While the visualisations do not necessarily fill these blanks, often they hint strongly at possible answers, perhaps providing directions for further research. For example, the sudden cessation of a stream of correspondence may not definitively indicate a person's death, but it is certainly suggestive. Similarly where a sender or recipient is ambiguous, the larger picture may well suggest an identity.

To illustrate how visualisations might throw light on the above areas, we can consider the case of William Montgomery. The migrant letter collections we looked at often feature letters sent between family members and names are often ambiguous. The numbers of letters attributed to individuals may appear to be relatively small, but this may be a result of this ambiguity. In the collections studied for this project there are entries for 'Montgomery, William', 'Montgomery, Wm', 'Montgomery, Wm.' and 'Montgomery, W.'. If we use a Bubble graph to chart the locations of these entries over time, we notice a great deal of contiguity in terms of place and time. For instance,'Montgomery, Wm' and 'Montgomery, W.' were both in Cincinnati around 1847/48.

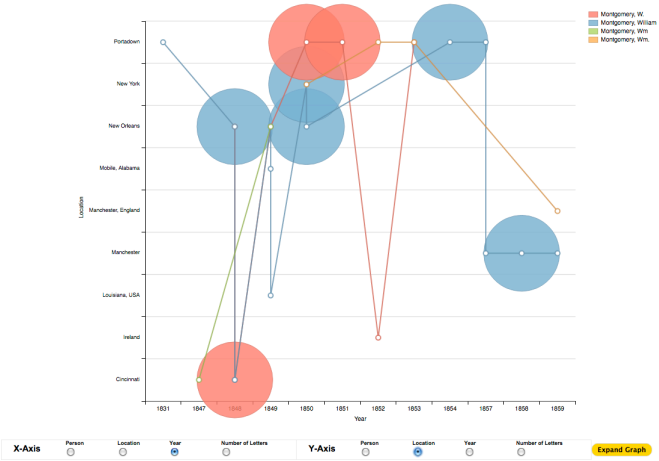

Figure 2. Bubble graph depicting movements of William Montgomery.

It is hardly conclusive, but a graph such as this gives us reason to suspect that these four variants represent the same person. If this is the case it would appear that William Montgomery travelled from Portadown to Cincinnati sometime between 1831 and 1847. He then travelled to New Orleans in 1848. With the exception of a trip to Mobile, Alabama in 1849, he seems to have lived in New Orleans until 1850, when he went to New York (perhaps to get the boat) and returned to Portadown. He remained in Portadown until 1857 when he went to Manchester, England.

More Info	Sender	Recipient	Origin	Destination	Date	Year
More info	Montgomery, William		Portadown		1831-01-16	1831
More info	Montgomery, Wm		Cincinnati		1847-11-23	1847
More info	Montgomery, William		Cincinnati		1848-01-11	1848
More info	Montgomery, W.		Cincinnati		1848-05-30	1848
More info	Montgomery, W.		Cincinnati		1848-09-21	1848
More info	Montgomery, William		New Orleans		1848-12-11	1848
More info	Montgomery, W.		New Orleans		1848-12-30	1848
More info	Montgomery, William		New Orleans		1848-12-30	1848
More info	Montgomery, Wm		New Orleans		1849-04-16	1849
More info	Montgomery, W.		New Orleans		1849-06-28	1849
More info	Montgomery, William		Mobile, Alabama		1849-09-07	1849
More info	Montgomery, William		Louisiana, USA		1849-10-06	1849
More info	Montgomery, William		New Orleans		1849-12-05	1849
More info	Montgomery, William		New Orleans		1850-03-13	1850
More info	Montgomery, William		New Orleans		1850-04-11	1850
More info	Montgomery, Wm.		New York		1850-06-04	1850
More info	Montgomery, William		New York		1850-06-06	1850
More info	Montgomery, William		New York		1850-06-11	1850
More info	Montgomery, W.		Portadown		1850-07-24	1850
More info	Montgomery, W.		Portadown		1850-11-06	1850
More info	Montgomery, W.		Portadown		1851-03-27	1851
More info	Montgomery, W.		Portadown		1851-09-25	1851
More info	Montgomery, W.		Ireland		1852-08-01	1852

Figure 3. Tabular display of Montgomery locations.

Of course, much of this may appear conjecture, but we do know that letters were sent from each of these locations on specific dates and a tabular listing ordered by date seems to bear out this narrative. For someone researching 'William Montgomery' of Portadown, this may give a springboard for further in-depth research - research that might corroborate the initial scenario. An analysis of the recipients of letters from these 'people' might also bolster the case. What is important is not that the details are relatively inconclusive, but that we can get at them so easily.

From a broader perspective, data visualisations form a context within which the individual narratives gain definition. The large movement of people to countries such as the United States can be seen as particular journeys starting in specific locales and progressing in idiosyncratic ways. The more comprehensive the dataset, the more detailed these narratives become. The problem with all of these analyses, however, is that we are rarely dealing with a comprehensive corpus. Taking the Smith/O'Gowan letters as an example, the larger part of the collection is a series of letters from Hugh Smith in New York to James Smith in Fall River, Massachusetts. There are an additional three letters from Ireland to James Smith. On the surface, James Smith would seem to be a privileged son with whom everyone was anxious to communicate, but who did not respond at all. Indeed Hugh Smith for all his endeavours does not receive any correspondence in return. However,

this collection is comprised of just 18 letters, so it is almost certain that not all of the letters are at our disposal. This trivial and obvious example points to larger issues regarding any data visualisation of large numbers of letters; we rarely, if ever, have all the data at our disposal and must bear in mind these gaps in our records when drawing conclusions. Nevertheless, from Hugh's correspondence to James, we know that Hugh was in New York and James was more than likely in Fall River in the period between 1887 and 1899. It might go without saying, but it is a fact that Hugh was alive when he wrote those letters and so was James, so we have at least some biographical data that helps create a picture of their lives.

A useful starting point when visualising metadata relating to migrant letter collections is to track the correspondents' movements over time. Figure 4 details the origin points (shown with red dots) and the destination points (shown in blue dots) for the letters contained in the four datasets. Unsurprisingly, given the datasets being used here, the majority of correspondents communicate between Ireland and various locations in America.

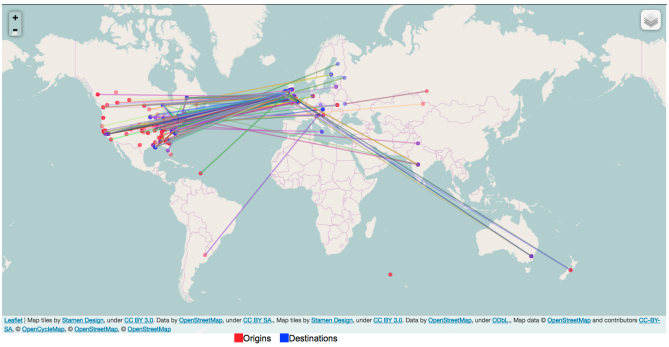

Figure 4. Distribution of correspondence for the IED, Lough, Smith/O'Gowan and DIL collections.

We can also explore the migratory patterns of an individual or family. The Lough collection, for example, is comprised of 99 letters. Of these letters we are missing the origins of 9 letters and the destinations of 12. In other words, we are missing 21 out of 198 location elements; all other locations are available and can be mapped with varying levels of granularity. Some places are as specific as 'Westfield, Hampden County, Massachusetts, America', while others are as general as 'Ireland' (Figure 5).

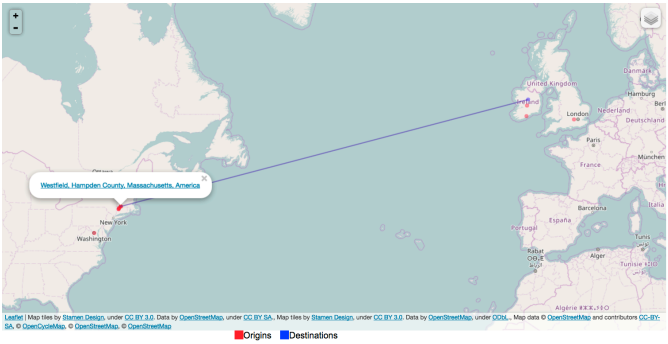

Figure 5. Map of correspondence of Lough collection.

Location data can be mapped, but other than telling us where letters were sent from and received, there is little to be gleaned. Arguably, a map only really becomes useful when another dimension is added. Interactivity is a powerful component that allows the user to choose what letters they want represented on the map using facets such as year, person or place (Figure 6), making it possible to narrow down a search to letters by a particular author, from a particular period, sent to or from a particular location. Additionally, a map that represents time graphically (such as that shown in Figure 7) helps us to understand where places figure in the temporal narrative of a collection. For example, Figure 7 demonstrates a consistent presence in Winsted, Connecticut, in the Lough letters, with later excursions into Westfield and then Torrington.

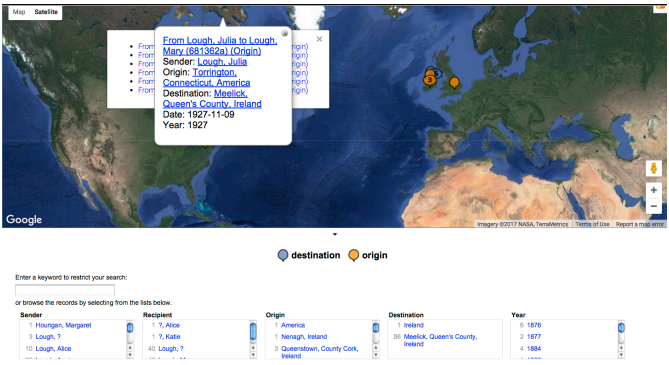

Figure 6. Map with facetted interactivity.

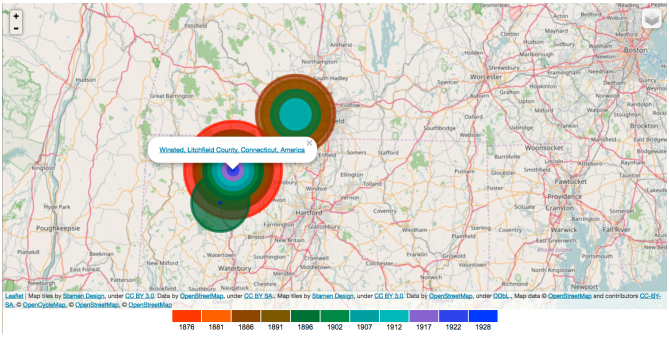

Figure 7. Time map.

The Bubble Graph represented in Figure 8 uses location and year as its axes. The bubble size represents the number of letters, while the colour represents the person. Unfortunately the large number of letters without a date in the Lough collection - 35 out of 99 - means we have a large number of letters without a year.

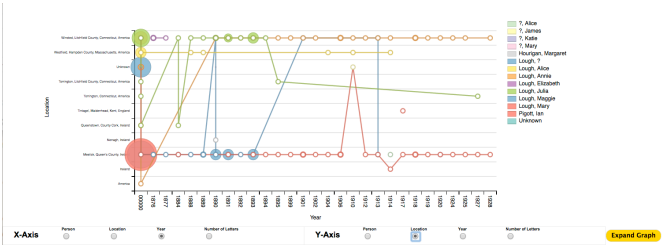

Figure 8. Bubble graph for the Lough collection.

As shown in the time map for the Lough collection (Figure 7), there is a persistent sample of letters originating or terminating in Winsted starting in 1876 and ending in 1928. Looking at Figure 8, we can see that Winsted is the earliest American location, featuring in letters written by Elizabeth, Annie, and Julia Lough. Elizabeth's letters are the earliest but they end (as far as this collection is concerned) in 1877. Julia begins writing from Winsted in 1884 and continues to do so until 1894. She takes up her pen again in 1895 but this time writing from Torrington. She writes once more from this location in 1927 (again as far as the collection reveals). Annie, however, begins writing from Winsted in 1890 and continues to do so until 1928, while Alice begins writing from Westfield in 1888 and continues to do so until 1914. In essence, Elizabeth and Annie do not move from Winsted once there, Alice seems to reside only in Westfield. Only Julia moves location, beginning in Winsted, but ultimately moving to and staying in Torrington.

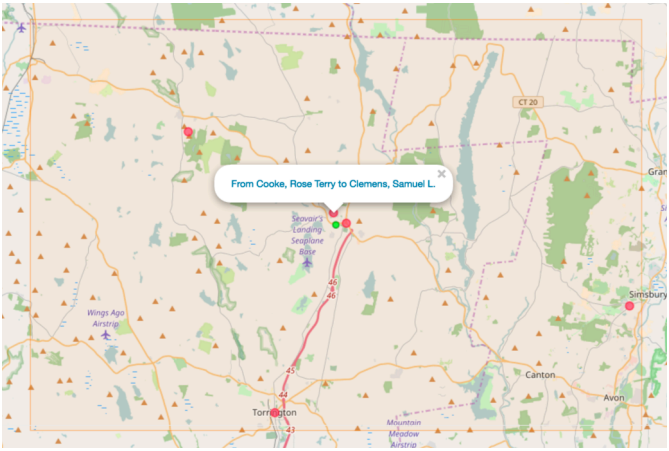

Figure 9. Correspondents also in the Winsted area.

When studying migration letters, the obvious approach is to restrict analyses to collections that are concerned only with the letters of migrants. However, in doing this one may miss unexpected connections that might throw more light on a correspondent. For example, we can also examine Winsted as a location that features in other letter collections. While not explicitly adding to our knowledge of the Loughs' own story, it does put their situation in context. For instance, also living in Winsted at the time of the Lough sisters was the American author, Rose Terry Cooke (17/02/1827 - 18/07/1892), a correspondent of Samuel Clemens (Mark Twain), who also visited the area (Figure 9). It is quite likely that the Lough sisters would at least have been aware of a celebrity in their midst. Additionally, in Fall River, three years after Hugh Smith's last letter to James Smith, James Whipp was writing to the Irish Socialist and revolutionary, James Connolly. If nothing else, this shows that Fall River was not a political backwater.

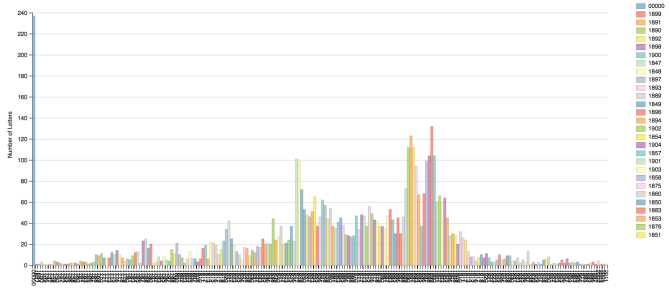

Figure 10. Concentration of letters.

So far our discussion has primarily focused on the Lough and Smith/O'Gowan collections. Once we add the IED and DIL collections we can start to appreciate general trends in the entire dataset. For instance, most of the letters seem to be from 1850 to 1900, with a majority of those from the 1890's (Figure 10). Looking at the correspondents themselves (Figure 11) 'Smyth, James A.' and 'Smyth, J. A.' are the top correspondents, meaning they appear as either sender or recipient in the most number of letters (305 and 237 respectively). In the absence of comprehensive personography information only closer analysis of the letters themselves can determine if these two correspondents are indeed one and the same person, in which case their proportionate representation in the dataset is very large. Indeed the need for more biographical data is particularly acute when one examines social networks. If we look at the principal players in the Lough and Smith/O'Gowan collections in the context of the wider dataset we find some new connections for James and Hugh Smith (see Figure 12). However, on closer inspection of the letters involved, there would appear to be two 'James Smith's and two 'Hugh Smith's, each from a different timeframe and location. The importance of disambiguating correspondents is a large and challenging task, but ultimately crucial. A graph such as that in Figure 12 can only suggest possible avenues of investigation, it cannot guarantee that they will bear fruit. However, having the metadata databased and available to query makes establishing dead ends a fairly quick process.

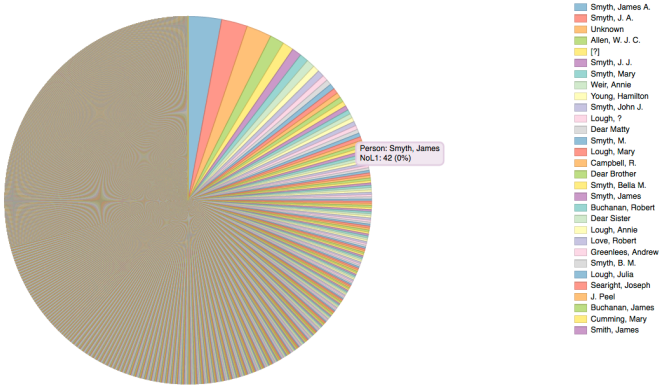

Figure 11. Number of letters by correspondent.

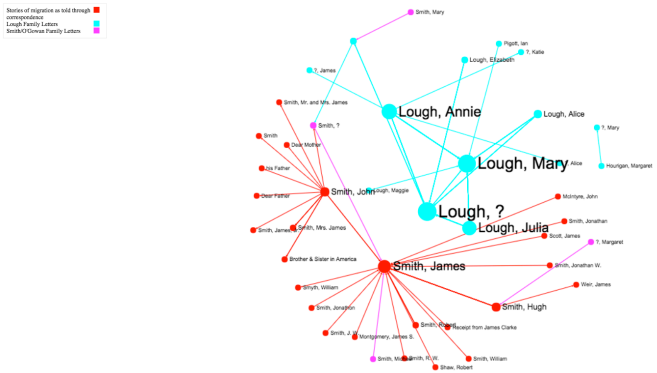

Figure 12. Social network graph.

Depending on the research question, the choice of graph or analysis used is important. A treemap or social network graph can be confusing when applied to a very large dataset, but is very useful when one can control the parameters and examine a subset of the data. Maps can give an overview of general trends, but often a bar or pie chart can make these trends even clearer. When examining an individual's or family's movements over time, bubble and line graphs can elucidate several vectors at once. Regardless of the technology employed or even the dataset used, a clearly articulated research question will often minimise time and suggest the most appropriate tool. Having said that, occasionally simply browsing the data using a variety of tools can highlight trends and outliers worthy of investigation.

Conclusion

What this article has hopefully highlighted is the potential for working with metadata such as that information embedded within the <correspDesc> element of the TEI header. If the same information categories are applied across letter collections it becomes easier for resources to interconnect, allowing larger data sets to be explored and compared using social network analysis techniques.[20] Work-

[20] Other projects which use TEI to capture personography and placeography information include: *Map of Early and Modern London*. Available from: https://mapoflondon.uvic.ca and https://mapoflondon.uvic.ca/historical_personography.xml; *Colonial Despatches: The colonial despatches of Vancouver Island and British Columbia 1846-1871*. Available from: http://bcgenesis.uvic.ca/places.xml; and *UCLA Encyclopedia of Egyptology*. Available from: https://uee.ats.ucla.edu/welcome/. Accessed March 15, 2017.

ing with five basic information fields (sender, recipient, origin, destination and date) is certainly a good starting point; however, capturing more detailed metadata relating to the participants and their locations opens up all sorts of possibilities for social network analysis, allowing the user to explore a migrant's life story (whether and when they married, whether they had children, their occupations, when they migrated and the passage they took etc.) and to make connections between patterns of migration and other factors such as age, sex, faith, educational background and social status. Provided one takes a well-structured approach to the basic metadata, this additional information can be integrated at a later point.

Subsequent to the DEM project, *Visual Correspondence* was developed to take a more general and expansive approach to visualising correspondence metadata. Instead of dealing exclusively with migrant letter collections, it uses correspondence of all varieties. Learning from the DEM project, a more standardised process for ingesting letter collections was developed for this site, making it easier to incorporate new collections. The site now contains over 164,000 letters from 53 collections. Collections are stored as distinct datasets, giving the user the power to create pseudo collections comprised of several datasets, while maintaining the integrity of the collections themselves. This means that unlike the initial DEM prototype, which amalgamated all the content as one collection, individual collections can now be analysed separately. The system provides a more sophisticated set of facets that makes it more flexible when creating complex queries based on people, places or dates. Linked Open Data is incorporated to provide biographical information on some of the more well-known letter writers included, such as Einstein and Joyce. In addition, by extending the scope to include letters of all types, it can contextualise specific letters, as was seen in the case of the Lough sisters in Winsted, and increase the opportunities for serendipitous discovery. From this perspective, *Visual Correspondence* demonstrates some of the possibilities inherent in all letter collections.